DOCUMENT RESUME

ED 443 875 TM 031 520

AUTHOR Irwin, Patrick M.; Plake, Barbara S.; Impara, James C.
TITLE Validity of Item Performance Estimates from an Angoff Standard Setting Study.
PUB DATE 2000-04-00
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Judges; Licensing Examinations (Professions); *Performance Factors; *Test Items; Validity
IDENTIFIERS *Angoff Methods; *Standard Setting

ABSTRACT

Judgmental standard setting methods, such as the W. H. Angoff (1971) method, use item performance estimates as the basis for determining the minimum passing score (MPS). Therefore, the accuracy of these item performance estimates is crucial to the validity of the resulting MPS. Recent researchers, (L. A. Shephard 1994; J. Impara, 1997) have called into question the ability of the judges to make accurate item performance estimates for the target subgroups of candidates, such as minimally competent candidates (MCCs). The purpose of this study was to examine the validity of the judges' estimates of MCC performance that are used to determine a minimum passing score in a standard setting study context. Results from the operational standard setting workshops for a certification program in financial management that used 29 judges in 1996 and 30 in 1997 provide evidence that item performance estimates were valid. Factors that might have influenced this high degree of validity in the item performance estimates in the standard setting study are discussed. (Contains 10 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 443 875

Validity of Item Performance Estimates from an Angoff Standard Setting Study

Patrick M. Irwin

University of Nebraska – Lincoln

Barbara S. Plake

University of Nebraska – Lincoln

James C. Impara

University of Nebraska – Lincoln

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Impara

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Author Contact:

Patrick M. Irwin

Millard Public Schools

5606 SO. 147th Street

Omaha, NE 68137-2604

Email: pirwin@millard.esu3.k12.ne.us

TM031520

BEST COPY AVAILABLE

Abstract

Judgmental standard setting methods, such as the Angoff (1971) method, use item performance estimates as the basis for determining the minimum passing score (MPS). Therefore the accuracy of these item performance estimates is crucial to the validity of the resulting MPS. Recent researchers (Shepard, 1994; Impara, 1997) have called into question the ability of the judges to make accurate item performance estimates for the target subgroups of candidates, such as minimally competent candidates (MCCs). The purpose of this study is to examine the validity of the judges' estimates of MCC performance that are used to determine a minimum passing score in a standard setting study context. Results provide evidence that item performance estimates were valid. Factors that might have influenced this high degree of validity in the item performance estimates in the standard setting study are discussed.

Validity of Item Performance Estimates from an Angoff Standard Setting Study

The purpose of this study was to examine the validity of the judges' estimates of MCC performance that are used to determine a minimum passing score in a standard setting study context. This study examined judges' ability to estimate item level group performance for the MCCs. The judges' estimates were compared to the actual performance of empirically-defined MCCs (candidates whose scores fell within 1 standard error of measurement of the MPS or cut score). For each item, actual performance data of MCCs was compared to the judges' estimated proportion of MCCs correctly answering the question. By comparing the judges' predictions of item performance with the item performance of candidates whose test scores identify them as minimally competent, the accuracy, or validity, of the judges' item performance estimates was investigated (Kane, 1994).

There is a large body of literature regarding various aspects of the Angoff (1971) standard setting method. This section focuses on literature relevant to the judges' ability to accurately predict item performance estimates for the MCCs. One of the main concerns regarding the Angoff standard setting method is the ability of the judges to conceptualize the MCCs and to predict their item performance. According to Shepard (1995), who reported an application of the modified Angoff method to determine cutscores for several achievement levels on an assessment, the panel of judges was able to distinguish between hard and easy items but were unable to correctly estimate item difficulty. Judges tended to under estimate the difficulty of hard items and under estimate the easiness of easy items and therefore set cut scores that are too extreme. The outcome of judges systematically underestimating the difficulty of hard items is to set a cutscore that is too high. Therefore, for a test containing mostly hard items, the resulting cutscore would be too high. The opposite is true with easy items; judges were found to under

estimate the easiness of easy items, indicating fewer candidates getting the item right then actually did. This would have the impact of setting a cutscore that was too low. Shepard hypothesized that the judges could not hold the hypothetical MCC in mind and make accurate estimates of item difficulty. Shepard's conclusion from this study was that the Angoff procedure is fundamentally flawed.

Lorge and Kruglov (1953) studied the effect of performance data provided to judges when estimating item difficulty. The groups differed in the amount of information given before making these estimates. One group was given item performance information for some of the test items and was asked to provide performance estimates on the remaining questions. The results of this study found that each group of judges could distinguish between difficult and easy items, but the groups differed on their student performance estimates. The two groups significantly underestimated the difficulty of hard items and believed the items to be easier then they actually were. The results of this study are consistent with Shepard's findings. The group of judges that was given the additional information predicted estimates closer to the true performance of the students. The group of judges that was not provided this information tended to underestimate or overestimate the performance of the students on the items.

There are numerous articles which discuss issues related to increasing the accuracy of the results when using the Angoff method in standard setting studies. According to Jaeger (1991), judges must be experts in the domain being assessed in the standard setting. Jaeger describes experts as possessing the following characteristics: (a) experts excel mainly in their own domain, (b) experts perceive large meaningful patterns in their domain of expertise, (c) experts perform rapidly in their domain of expertise, (d) experts see and represent a problem in their domain at a deeper level than do novices, (e) experts spend time analyzing a problem qualitatively, (f)

experts have strong self monitoring skills, (g) experts are more accurate than novices at judging problem difficulty and (h) expertise lies more in an elaborated semantic memory than in a general reasoning process. The number of judges sampled should be sufficiently large to minimize error and also provide precise estimations of the standard that would represent the entire population of judges.

Mills, Melican and Ahluwalia (1991) addressed the importance of the selection process and training of the judges. These authors stress the importance of providing a clear definition of the concept of the MCC and selecting judges who are knowledgeable about the test materials, the candidate population, and the methodology used in the standard setting study. This training is crucial to reaching accurate results in standard setting.

Plake, Melican and Mills (1991) discuss the factors influencing intrajudge consistency during standard-settings. This article discusses various strategies to improve intrajudge consistency. The authors recommend extensive training of the judges, a clear definition of the knowledge, skills, and abilities (KSAs) of the MCCs, providing the answer keys to the judges, providing the judges with item performance data, and giving the judges the option to change their estimates after they are given performance data.

In order to provide evidence to support the validity of cutscores, Kane (1994) recommended various validation strategies related to both the procedures used and the results of those procedures. Among these strategies those that have a direct impact on the judges' ability to estimate item performance: (a) defining the goals of the standard setting study, (b) selecting qualified judges, (c) providing proper training of judges, (d) providing performance data to judges and (e) defining performance standards. Kane discusses two methods that use the data from a standard setting study to document the validity of the outcome of the process

(i.e., cutscore). The first method compares the judges' item performance estimates to the actual proportion of MCCs who correctly answer the item. This procedure provides some evidence that the judges are able to conceptualize the MCCs and are able to predict their item performance. The second method involves identifying two groups and making performance comparisons. One group consists of candidates with scores that were just barely passing and the other group consists of candidates who just barely failed. The two groups would be compared on the proportion that correctly answered the items. The higher scoring group should outperform the lower group on each of the test items. These results also add support to the validity of the cutscore.

In this study, one of Kane's empirical methods of providing evidence for the cutscore was used. By selecting out those candidates who score close to the MPS, a group of empirically-defined MCCs was formed. Their item performance (proportion correctly answering the item) was compared to the judges' average estimate of MCC item performance. The magnitude of this difference has direct implications for the validity the MPS based on the judges' item performance estimates.

Method

This study examined the validity of the judges' ability to accurately determine a minimum passing score in a standard setting study context. The study used a validation method suggested by Kane (1994) of comparing the proportion of MCCs who correctly answer an item to the average item performance estimate from the panel of judges.

Instruments

The data for this study was taken from operational standard setting workshops for an international certification program in financial management. The test consists of 230 multiple-

BEST COPY AVAILABLE

choice questions designed to measure entry-level knowledge. The data were gathered over the 2 year period from 1996 to 1997. Each year a new standard setting study was conducted to determine the cutscore for that year's test. Every year different 230-item exams were built to the same content specifications. The test was developed by the certification program to measure the eight content domains described in the table of specifications. Table 1 shows descriptive statistics from the 1996 and 1997 examinations, including total number of candidates, overall test means and standard deviations, and the internal consistency reliability estimates.

Table 1. Descriptive Statistics for the Examination

Year of Test	N	Mean (SD)	Reliability (KR 20)
1996	14,381	144.35 (30.36)	0.9544
1997	16,832	146.37 (32.03)	0.9574

From the 230 test questions, two 115-item psychometrically equivalent forms were prepared (forms A and B). These forms were designed to be as equivalent as possible in terms of: (a) coverage of the table of specifications for the certification examination, (b) average difficulty within content domain, (c) overall average performance, (d) and internal consistency reliability.

Judges

In 1996 29 judges were convened to participate in a standard setting workshop. In 1997 a total of 30 judges participated. Eleven judges on the 1996 panel had participated in a previous standard setting study and 18 judges were unique to the procedures. In 1997 there were 15 repeat judges from 1996 and 15 judges who had no prior standard setting experience. Each year the panel of judges was selected by the organization to be representative of the demographics of the

organization. Each year the judges were divided into two groups, A and B. The groups were divided in such a way as to maintain equal representation.

Standard Setting Study Procedures

The procedures in both years were the same. Judges met for a 2-day standard setting study. The two groups of judges (A and B) were kept together, and trained as a group, to maximize consistency in training across groups. The training consisted of an explanation of the standard setting study, a discussion that was designed to elicit the KSAs of the MCC, a review of the table of specifications for the test, and experience on a set of practice items. By allowing the judges to familiarize themselves with these procedures it was thought they would have a better understanding of the standard setting process. In the practice session, judges made item performance estimates on practice items (items administered in a proir year). This was done to explain the process and to address any questions or concerns. After practice, the judges were divided into their designated groups, given a copy of the test form assigned to their group (either form A or B) and asked to make independent item performance estimates for each item. Judges' Round 1 estimates and materials were collected.

At the start of day 2, judges were given item performance data from the most recent administration of the examination. This included information about the proportion of candidates who would pass using the cutscore based on the Round 1 item performance estimates on their 115-item test form. Next the judges completed Round 2 where they were given the opportunity to make adjustments to their item performance estimates made on Round 1.

The 1997 examination was built to the exact same specifications as the 1996 examination. The 1997 panel of judges was selected in the same fashion as the 1996 panel and followed the same training procedures. As before, the panel of judges was divided into two

BEST COPY AVAILABLE

equally representative groups and the 230 item test was divided into two psychometrically equivalent 115 item tests (Form A and B).

Cutscores

The standard setting studies were performed following the procedures stated above and the judges' individual item performance estimates were averaged for each group to identify a cutscore for each group. Each group examined half (115) of the total (230) test questions. In addition, both groups made item performance estimates on a common set of 48 items (known as Form C). Form C consisted of 24 items drawn from Form A and 24 drawn from Form B. The ratings on Form C were compared across groups to ensure the group ratings were on the same scale. Each group made item performance estimates (both at Round 1 and Round 2) for the 48 items after they had completed their estimates for their respective form. Because the group ratings on Form C differed less than would be expected due to chance, the two cutscores for Forms A and B were added together to produce one cutscore for the 230-item test.

Results from the standard setting workshops, including the judges' MPS and variability, were communicated to the organization's Board of Trustees. They set the final operational MPS, based in part on the results of the standard setting workshops, but also considered other policies and related issues. These procedures were followed in both 1996 and 1997. The panel of judges in the 1996 standard setting study arrived at a cutscore of 148.46; the Board set the cutscore at 146. The panel in 1997 set the cutscore at 148.54 and the Board set the score at 148. Table 2 shows the judges' cutscores and the Board's cutscores.

Table 2. MPS Recommended by the Panel and the Board's MPS Values

---

Year of Test	MPS (Judges)	MPS (Board)
1996	148.46	146
1997	148.54	148

Candidates

The total number of candidates who participated in the examinations was 14,381 in 1996 and 16,832 in 1997. From this pool of candidates this study considered those candidates who were identified as the "Empirically Minimally Competent Candidates" (EMCCs). These candidates had total scores on the examination that fell within one standard error of measurement from the Board identified cutscore.

Analytic Procedures

The data from the 1996 and 1997 examinations were used to calculate the standard error of measurement. This value was used to calculate the 67% confidence interval around the Board specified MPS. All candidates whose test scores for that year fell within this interval were identified as EMCCs. The EMCCs were then selected out of the data set. EMCC item performances were calculated for each of the 230 questions for each of the examinations. This item performance information was compared to the judges' item performance estimates. Therefore, for each question on the examination, the EMCC's performance data were compared to judges' average item performance estimate. This information allowed for an examination of how closely the judges' estimated the EMCC actual performance. These procedures were followed for 1996 and then repeated for the 1997 data.

Analysis

In 1996 the cutscore was set at 146 and the standard error of measurement was 6.48, raised to the nearest integer, 7. The candidates whose test scores fell between 139 and 153 for the 1996 examination were identified as EMCCs. There were 2,731 candidates identified as EMCCs out of 14,381 candidates who took the examination in 1996. In 1997 the cutscore was set at 148 and the standard error of measurement was 6.61 (rounded up to 7). The candidates whose test scores fell between 141 and 155 for the 1997 examination were identified as EMCCs. There were 2,960 candidates out of 16,832 that fell in to the classification of an EMCC. Table 3 provides the standard error of measurement, cutscores, and the number of EMCCs for each year.

Table 3. Identification of the Empirical Minimally Competent Candidates (EMCCs)

Year Test	MPS (cutscore)	SEM	+ or - 1 SEM	# of EMCCs	Total # Cand.
1996	146	6.48 = 7	139 - 153	2,731	14,381
1997	148	6.61 = 7	141 - 155	2,960	16,832

The item performance estimates made by the panel of judges were subtracted from the proportion of EMCCs answering an item correctly. These differences are displayed as differences and as absolute differences. The average differences alone could be misleading because the process of averaging these differences should balance out due to the combining of positive and negative differences. Also, because the EMCC group was selected in a symmetrical manner around the Board-imposed cutscore, and this cutscore was close to the average of the judges' individual MPSs, there is the expectation that these differences will average to near zero. The absolute difference eliminates the negative and positive signs so this is a more appropriate indicator of accuracy of the item performance estimates. The averages of the absolute differences

show how close in magnitude the judges' item performance estimates are to the actual performance of the EMCCs. Therefore, these measures answer the overarching question of this study, whether the judges are able to predict an accurate and valid estimate of the MCCs performance when determining a minimum passing score in a standard setting context.

As a follow up analysis, the magnitude of the actual and absolute difference was examined by item difficulty level. Previous research (Shepard, 1995 Impara & Plake; 1998) indicated that accuracy of item performance estimates might vary as a function of item difficulty. To examine this possibility, the magnitude of the differences in item performance estimates with actual EMCC item performance was examined for evidence of a pattern related to item difficulty. The proportion of items that yielded an absolute difference of 0.1000 or higher was arbitrarily set for selecting items for further consideration (Impara & Plake, 1998). The purpose of these follow up analyses was to provide additional substantive information relevant to judges' accuracy in making item performance estimates. Unfortunately, due to the secure nature of the test and the limited information about the items in the data set, specific item features such as item format, inclusion of graphics, or content category designations in the Table of Specifications, could not be examined.

Results

The results of this study are based on the judges' item performance estimates and the EMCC's actual item performance. For each test the actual performance of the EMCCs for all 230 questions and the judges' estimates for the 1996 and the 1997 exam data was determined. From this information, the judges' estimate was subtracted from the EMCCs performance. The results from the overall average of these differences indicate that there is substantial agreement between the judges' estimates of the MCCs and the actual performance of candidates whose scores are

BEST COPY AVAILABLE

close to the MPS. As anticipated, the average difference in judges' estimates and EMCC performance was near zero. When the absolute value of the differences was averaged, the result was .0752 in 1996 and .0730 in 1997. Table 4 summarizes the 1996 and 1997 results. Taken together, and across years, these results indicate that there was a very high degree of accuracy in the judges' anticipated performance of the MCC.

Table 4. Accuracy of Judge's Estimates

Year of Test	Average Difference Judges - MCCs	Average Absolute Difference
1996	-.0016	.0752
1997	-.0022	.0730

Items with absolute differences greater than 0.1000 were examined. For these items, the differences were ranked according to magnitude based on the EMCC's performance on the item. Items were categorized as hard, moderate, or easy. If the proportion of EMCCs getting the item correct was 0.3000 or lower, the item was considered a hard item. If the proportion was 0.7000 or higher it was considered an easy item; otherwise the item was classified as moderately difficult. In 1996 there were 68 items that had an absolute difference greater than 0.1000 and in 1997 there were 58 items. For 1996 there were fewer hard items (8) with differences over 0.1000 compared to easy items (25). The majority of items (35) were classified as moderately difficult in the 1996 data. In 1997 there were 9 hard items with differences over 0.1000 compared to easy items (26). The majority of items (23) were classified as moderately difficult.

Across years, twice as many easy items than hard items had differences greater than 0.1000. For these hard items, judges routinely over estimated how well the candidates would

perform. For these easy items, the results indicate the judges systematically underestimated the performance of these candidates. The judges also tended to over estimate candidates' performances on these moderately difficult items. In terms of degree of accuracy, the average difference for these hard items was -0.1689 in 1996 and -0.2046 in 1997. The average difference for these medium items was -0.1239 in 1996 and -0.1450 in 1997. The average difference for these easy items was 0.1249 in 1996 and 0.1379 in 1997. In terms of absolute value, these average absolute differences were 0.1689 in 1996 and 0.2046 in 1997. The average absolute difference for the medium items was 0.1309 in 1996 and 0.1450 in 1997. The average absolute difference for the easy items was 0.1249 in 1996 and 0.1379 in 1997.

The sign (or direction) of the difference was also a concern; if the sign of the difference was positive the judges tended to underestimate the difficulty and if the sign was negative they overestimated. The results for the positive and negative differences indicate a tendency of the judges to overestimate the candidates' performance (two thirds of the performance estimates were negative).

Discussion and Conclusion

This study addressed the validity of estimates provided by a panel of judges in setting a cutscore in an Angoff standard setting study. The results of this study indicate that the panels of judges used in this study were able to accurately identify and estimate the performance of the MCCs on an international certification examination in financial management.

The items that had an absolute difference of 0.1000 or greater when subtracting the judges' estimates from the MCCs performance were singled out for further study. Judges tended to overestimate candidate performance on hard and moderately difficult items, but underestimated candidate performance on easy items. Several differences between this study and

Shepard's 1995 study should be noted. First, Shepard did not limit her analysis to only the items with absolute differences greater than 0.1000. Second, Shepard's analyses were not based on comparing judges' estimates with the actual performance of empirically defined MCCs. These differences should be kept in mind when comparing the results of this study to those of Shepard.

Another point to consider when interpreting these results is the fact that the actual operational cutscore was used, rather then the judges' MPS, when identifying the EMCCs. However, the differences in magnitude with the Board-determined cutscore and the judges' MPS were minimal (1996 cutscore = 146, MPS = 148; 1997 cutscore = 148; MPS = 148).

A critically important issue in standard setting is the training of the judges. If the judges are trained properly and given careful guidance, they should be able to successfully complete the tasks at hand. As proposed in the literature (Melican, Mills & Ahluwaha, 1991; Reid, 1991; Cizek, 1996), the training should consist of a clear explanation of the standard setting study, a clear definition of the KSAs of the MCCs, a review of the table of specifications, and training on practice items. The amount of time spent on training may be a factor in increased validity of the results of the standard setting study.

In the standard setting procedures used for this study training took a central role. For each examination in each of the years, the panelists participated in an in-depth training session lasting approximately 4 hours. Over 1 hour was devoted to a discussion designed to elicit from the panelists the knowledge, skills, and abilities of the MCC, focusing specifically on the components of the table of specification for the examination. At the completion of the discussion, panelists were given copies of these KSAs listed by component of the table of specifications. As part of the evaluation process, panelists were asked to rate the quality of the training and the specific components. High ratings were uniformly provided by the panelists on

the "Discussion of the MCC", as well as the other training components.

The results of this study were limited due to the fact that the actual items were not available because of test security. The fact that several items that had a difference of 0.1000 or greater, when subtracting the judges' estimates from the MCCs performance, raised many questions. Those items with larger absolute differences might, for example, have come from only a few of the content categories from the Table of Specifications. Or, perhaps these items tended to be multi-step problems or items with multiple parts. These are only a few of the questions raised. The availability of the actual test questions might have explained some of these differences or revealed some other patterns.

This study only examined one cutscore; the study did not address a panel's ability to set more than one cutscore for the same assessment, as was the case in Shepard's study. Issues of reliability were not addressed in this study; however, reliability plays a very important part in the standard setting study.

Future research should focus on the generalizability of these results and the conditions that supported the validity evidence found in this study. Training has been highlighted as one possible link to the quality of these results. There are other factors, specific to this particular standard setting situation, which may have influenced the degree of accuracy in the judges' item performance estimate. The content is quantitative in nature, which may have facilitated the judges' grasp and use of the item performance data. The overall content is generally homogeneous even though it is broken down into eight content categories. The candidates are highly motivated to perform well, as employment options are directly linked to passing the test. Therefore, cautions should be exercised when generalizing the results to other standard setting contexts.

The purpose for a standard setting study is to make the most accurate prediction of a minimum passing score possible. Important decisions are based on these cutscores. Decisions will be made based on these scores that will affect the candidates' future. A great deal of time and money is spent on the procedures to set these passing scores. If the proper procedures are not followed and the results are not valid then the decisions that are based on these results do not accurately represent the candidates' true ability. The need to address the validity of the standard setting study and the cutscores is of great importance.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 508-60). Washington, DC.: American Council on Education.

Cizek, G. (1996). Standard setting guidelines. Educational Measurement: Issues and Practice, 15, 13-21, 12.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Journal of Educational Measurement, 35(1), 69-81.

Jaeger, R.M. (1991). Selection of judges for standard-setting. Educational Measurement: Issues and Practice, 10(2), 3-14.

Kane, M.T. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-461.

Lorge, I., & Kruglov, L.K. (1953). The improvement of the estimates of test difficulty. Educational and Psychological Measurement, 13, 34-46.

Mills, C.N., Melican, G.J., & Ahluwalia, N.T. (1991). Defining minimal competence. Educational Measurement: Issues and Practice, 10(2), 7-10.

Plake, B.S. (1998). Setting performance standards for professional licensure and certification. Applied Measurement in Education, 11, 65-80.

Plake, B.S., Melican, G.J., & Mills, C.N. (1991). Factors influencing intrajudge consistency during standard-setting. Educational Measurement: Issues and Practice, 10(2), 15-25.

Reid, J.B. (1991). Training judges to provide standard-setting data. Educational Measurement: Issues and Practice, 10(2), 11-14.

Shepard, L.A. (1994, October). Implications for standard setting of the NAE evaluation of NAPE achievement levels. Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, National Assessment Governing Board, National Center for Educational Statistics, Washington, DC.